TECHNICAL UNIVERSITY OF MOMBASA

FACULTY OF APPLIED AND HEALTH SCIENCES

DEPARTMENT OF MATHEMATICS & PHYSICS

UNIVERSITY EXAMINATION FOR:

THE DEGREE OF BACHELOR OF SCIENCE IN STATISTICS & COMPUTER SCIENCE

AMA 4320: STATISTICAL MODELLING

END OF SEMESTER EXAMINATION

SERIES:APRIL2016

TIME:2HOURS

DATE: Pick DateMay2016

Instructions to Candidates

You should have the following for this examination

-*Answer Booklet, examination pass and student ID*

This paper consists of **FIVE** questions. Attemptquestion ONE (Compulsory) and any other TWO questions.

Do not write on the question paper.

Question ONE

a) Below is an example of a loop statement in R;

```
q <- 5
while(q <=10) {
    q <- q+1
    print(q)   }
```

i) Identify the loop statement in the above code. (2 marks)

ii) List the output from the above loop code. (3 marks)

b) Use the scan function in R to read the following data; 1 2 3 4 5 6 7 8 9 0. (3 marks)

c) i) Write a R code that will generate 1000 normal distribution random numbers with a mean of zero and standard deviation of 1. Assign these random numbers to a vector called y. (3 marks)

ii) You first want to get a summary of vector y. Write code in R to generate the mean, median, standard deviation and Interquartile range. (2 marks)

iii) Plot a histogram of vector y and superimpose a normal curve. (5 marks)

d) Provided with a 2X2 matrix ‘c’, write a code to generate the inverse (C^{-1}) of the matrix. (2 marks)

e) You have been provided with a hypothetical class of 120 students and the following R code;

```
> sample (1:120,40)
```

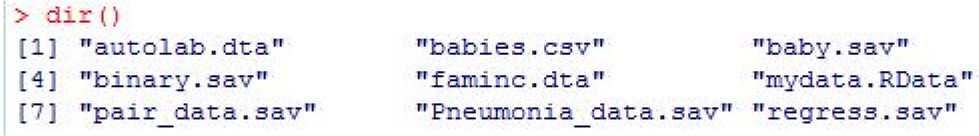
Explain the output of this code. (5 marks)

f) i) A researcher suggests the one month intensive exercises would reduce weight significantly. To prove his theory, he recruited 840 over-weight persons and subjected them to one month intensive exercises. The researcher recorded their weight before starting the exercises and after the exercises. What statistical test would the researcher use to show that one month intensive exercises reduces weight? (2 marks)

ii) Suppose the weight variable before starting the exercises was labelled ‘*weight0*’ and the weight variable after the exercises was labelled ‘*weight1*’. Write a code in R to test the null hypothesis of no association between one month intensive exercises and reduction in weight using the statistical test you named above. (3 marks)

Question TWO

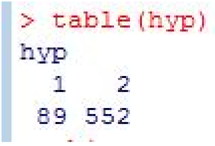
Suppose you are working on a directory that contains the following files;



```
> dir()
[1] "autolab.dta"        "babies.csv"         "baby.sav"
[4] "binary.sav"         "faminc.dta"         "mydata.RData"
[7] "pair_data.sav"      "Pneumonia_data.sav" "regress.sav"
```

i) Write a code in R that can read the "babies.csv" dataset. (2 marks)

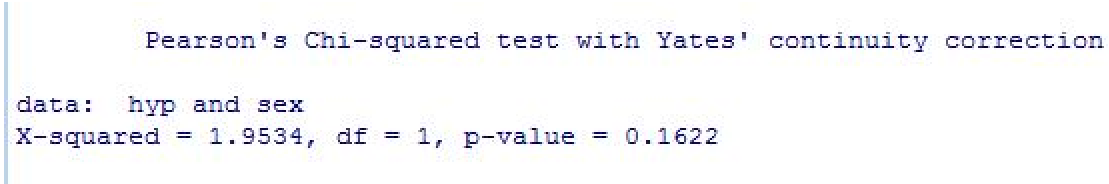
ii) The dataset contains two categorical variables; `*hyp*'-a binary variable coded Yes (1) or No (2) for presence or absence of hypertension respectively and `*sex*'-either a male or female. Explain and interpret the following R code on the dataset. (3 marks)



```
> table(hyp)
hyp
  1   2
 89 552
```

iii) To test the association between presence of hypertension and sex, you decide to run a chi-square test of association. Write a code to run the chi-square test in R. (5 marks)

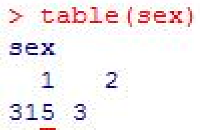
iv) After running the chi-square test above, you got the following results;



```
        Pearson's Chi-squared test with Yates' continuity correction

data:  hyp and sex
X-squared = 1.9534, df = 1, p-value = 0.1622
```

Give an interpretation of the results obtained. (5 marks)

v) Suppose to test if the assumption for using chi-square test, you run and obtained the following results;

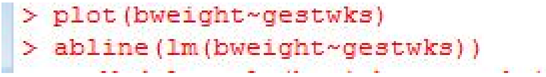


```
> table(sex)
sex
  1   2
315 3
```

Were the assumptions for using the chi-square test met? If no, suggest an alternative statistical test. (5 marks)

Question THREE

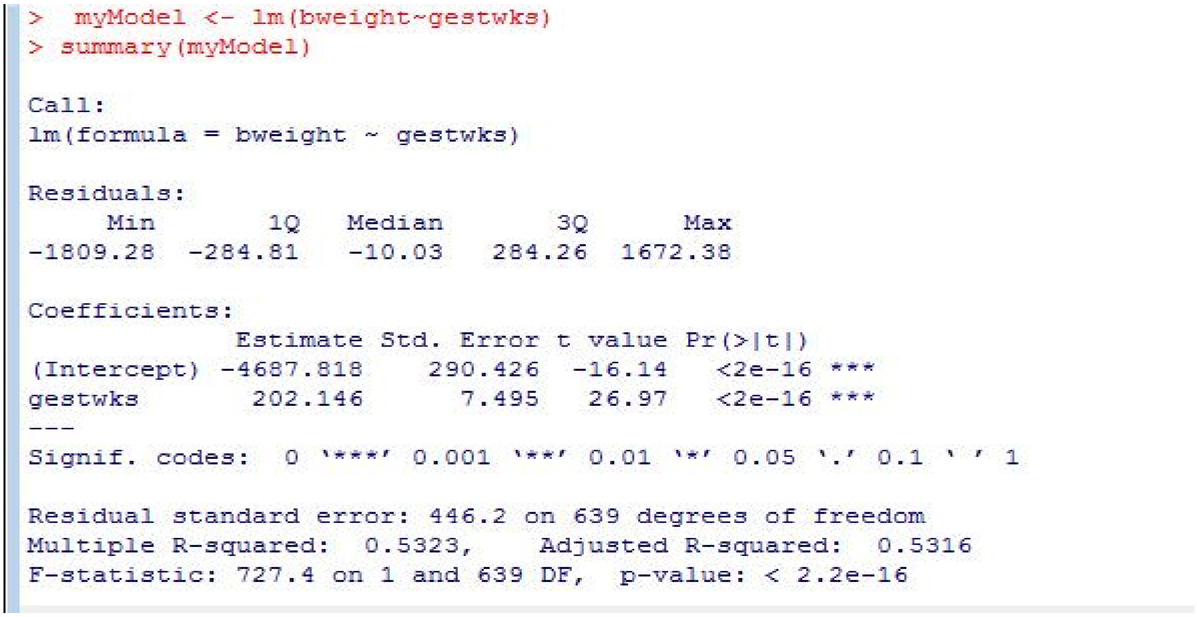
It is hypothesized that the birth weight of a baby is dependent on the gestation age in weeks. You have been hired to help analyze data collected to test the hypothesis; no association between birth weight and gestation age in weeks. You start by writing out the following R codes;



```
> plot(bweight~gestwks)
> abline(lm(bweight~gestwks))
```

i) Explain the results and reason for running the above script. (5 marks)

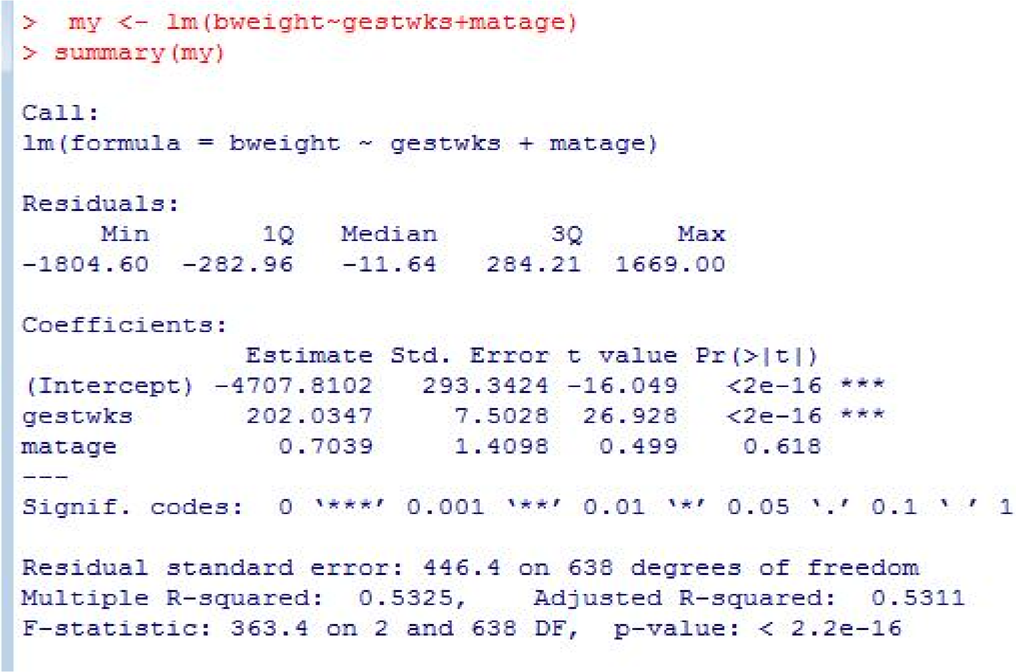
ii) Next you run the following R code and produce the results below;



```
>  myModel <- lm(bweight~gestwks)
> summary(myModel)

Call:
lm(formula = bweight ~ gestwks)

Residuals:
     Min       1Q   Median       3Q      Max 
-1809.28  -284.81   -10.03   284.26  1672.38 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4687.818    290.426  -16.14   <2e-16 ***
gestwks       202.146      7.495   26.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 446.2 on 639 degrees of freedom
Multiple R-squared:  0.5323,    Adjusted R-squared:  0.5316 
F-statistic: 727.4 on 1 and 639 DF,  p-value: < 2.2e-16
```

Write the linear regression model from the output above. (5marks)

iii) Next, the researcher suggest that the mother age could affect the birth weight and asks you to add the mother age to the regression model.



```
>  my <- lm(bweight~gestwks+matage)
> summary(my)

Call:
lm(formula = bweight ~ gestwks + matage)

Residuals:
     Min       1Q   Median       3Q      Max 
-1804.60  -282.96   -11.64   284.21  1669.00 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4707.8102   293.3424 -16.049   <2e-16 ***
gestwks       202.0347     7.5028  26.928   <2e-16 ***
matage          0.7039     1.4098   0.499    0.618    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 446.4 on 638 degrees of freedom
Multiple R-squared:  0.5325,    Adjusted R-squared:  0.5311 
F-statistic: 363.4 on 2 and 638 DF,  p-value: < 2.2e-16
```

Write the new regression model and explain if adding the variable has improved the models goodness of fit. (7 marks)

iv) What can you conclude from the analysis? (3 marks)

Question FOUR

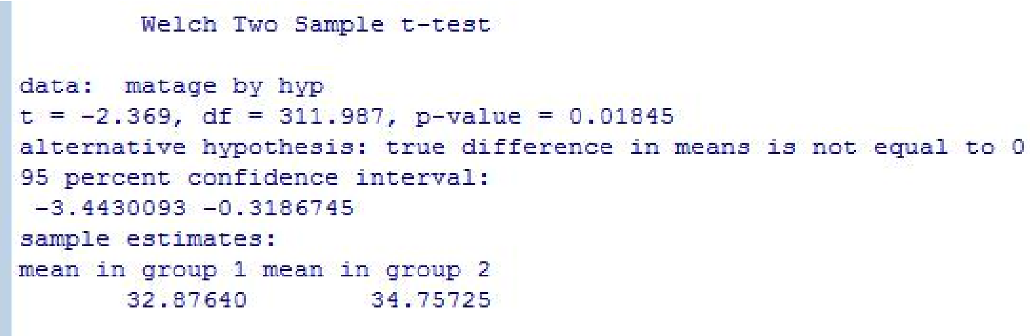
A researcher is interested in finding out whether age is associated with hypertension. He suggest that elderly people are likely to develop hypertension than younger ones. Using data from 641 people aged 18years and above, the researcher has asked you to help with the analysis. The variable for age in years is called ‘*matage*’ and is a continuous variable while the variable for hypertension is called ‘*hyp*’ and is a binary variable.

i) What is the appropriate statistical test to use to test the null hypothesis of no association of age in years with hypertension? (4 marks)

ii) State two assumptions you will be making in this analysis. (4 marks)

iii) Using the variables given, write a code in R to run the statistical test you listed above. (5 marks)

iv) Below are the output of the analysis;



```
        Welch Two Sample t-test

data:  matage by hyp
t = -2.369, df = 311.987, p-value = 0.01845
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.4430093 -0.3186745
sample estimates:
mean in group 1 mean in group 2
       32.87640        34.75725
```

State the mean difference in weight and the 95% confidence interval. (2 marks)

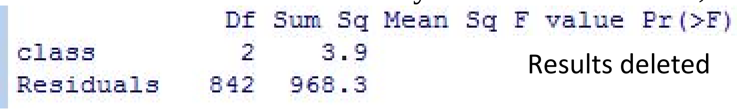
v) What can we conclude from these results? (5 marks)

Question FIVE

Maseno high school has three form IV classes that are taught mathematics by different teachers. A researcher is interested in testing if the mathematics performance across the three classes is the same, the three classes were offered the same exam at the end of term. Student mathematics score is a normally distributed continuous variable.

a) Suggest the appropriate statistical test to test the null hypothesis of no mean difference of mathematics score across the three classes. (2 marks)

b) Below are the results of the analysis the researcher did;



```
              Df Sum Sq Mean Sq F value Pr(>F)
class          2    3.9
Residuals    842  968.3
```

Results deleted

Some of the output has been erased. You are required to compute the mean Sq for both class and Residuals and the F value. (9 marks)

c) If the dependent variable (mathematics score) was not normally distributed, what other appropriate statistical test would you use? (2 marks)

d) Interpret the output. What conclusion can you make from the results? (7 marks)